NATURAL HISTORY MUSEUM

No specimen left behind:

Collections digitisation at the NHM, London*

Vince Smith

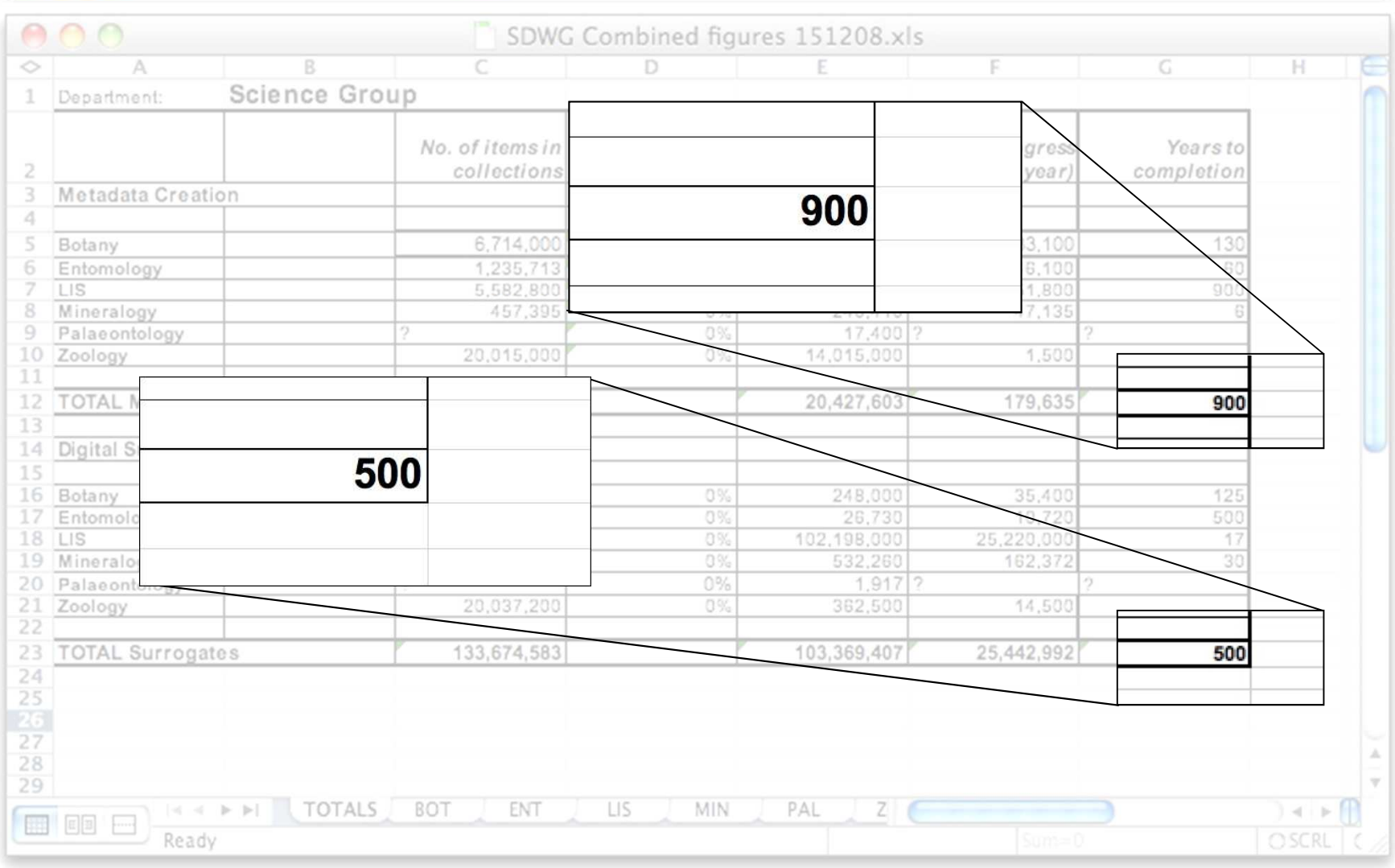
Collections for the 21st Century, Florida
5-6 May 2014

Some history…

"the rate of progress by the UK taxonomic institutions in digitising and making collections information available is disappointingly low… there is a significant risk of damage to the international reputation of major institutions such as The Natural History Museum"

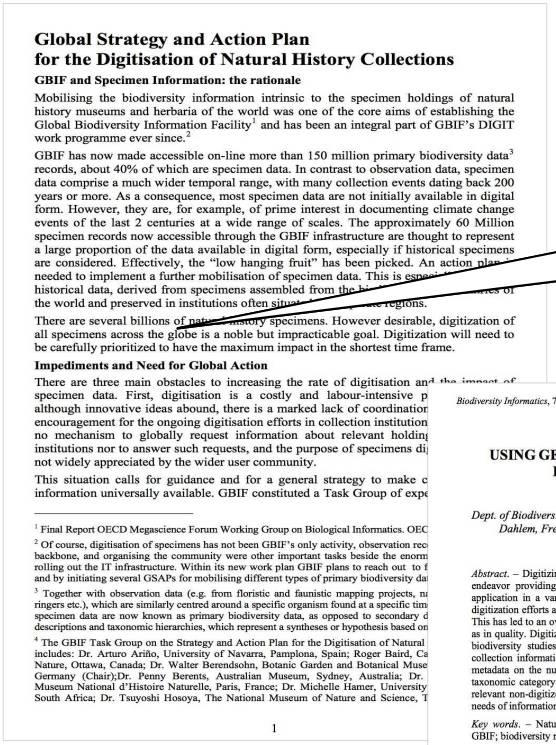
House of Lords Science and Technology Committee
Report on Taxonomy and Systematics, 2009

Digitisation rates at the NHM (circa 2009)



900 years to digitise the collection!

The prevailing attitude collections digitisation



Global Strategy and Action Plan for the Digitisation of Natural History Collections

GBIF and Specimen Information: the rationale

Mobilising the biodiversity information intrinsic to the specimen holdings of natural history museums and herbaria of the world was one of the core aims of establishing the Global Biodiversity Information Facility[1] and has been an integral part of GBIF's DIGIT work programme ever since.[2]

GBIF has now made accessible on-line more than 150 million primary biodiversity data[3] records, about 40% of which are specimen data. In contrast to observation data, specimen data comprise a much wider temporal range, with many collection events dating back 200 years or more. As a consequence, most specimen data are not initially available in digital form. However, they are, for example, of prime interest in documenting climate change events of the last 2 centuries at a wide range of scales. The approximately 60 Million specimen records now accessible through the GBIF infrastructure are thought to represent a large proportion of the data available in digital form, especially if historical specimens are considered. Effectively, the "low hanging fruit" has been picked. An action plan
needed to implement a further mobilisation of specimen data. This is
historical data, derived from specimens assembled from the
the world and preserved in institutions often situ
regions.

There are several billions of
all specimens across the globe is a noble but impracticable goal. Digitization will need to be carefully prioritized to have the maximum impact in the shortest time frame.

Impediments and Need for Global Action

There are three main obstacles to increasing the rate of digitisation and
specimen data. First, digitisation is a costly and labour-intensive p
although innovative ideas abound, there is a marked lack of coordination
encouragement for the ongoing digitisation efforts in collection institution
no mechanism to globally request information about relevant holding
institutions nor to answer such requests, and the purpose of specimens di
not widely appreciated by the wider user community.

This situation calls for guidance and for a general strategy to make c
information universally available. GBIF constituted a Task Group of expe

[1] Final Report OECD Megascience Forum Working Group on Biological Informatics. OEC

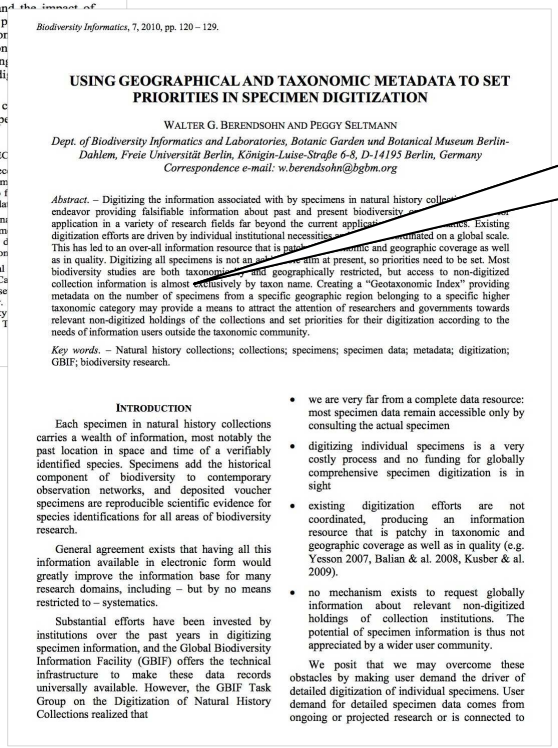
[2] Of course, digitisation of specimens has not been GBIF's only activity, observation rec
backbone, and organising the community were other important tasks beside the enorm
rolling out the IT infrastructure. Within its new work plan GBIF plans to reach out to f
and by initiating several GSAPs for mobilising different types of primary biodiversity da

[3] Together with observation data (e.g. from floristic and faunistic mapping projects, n
ringers etc.), which are similarly centred around a specific organism found at a specific tim
specimen data are now known as primary biodiversity data, as opposed to secondary d
descriptions and taxonomic hierarchies, which represent a syntheses or hypothesis based on

[4] The GBIF Task Group on the Strategy and Action Plan for the Digitisation of Natural
includes: Dr. Arturo Ariño, University of Navarra, Pamplona, Spain; Roger Baird, Ca
Nature, Ottawa, Canada; Dr. Walter Berendsohn, Botanic Garden and Botanical Muse
Germany (Chair);Dr. Penny Berents, Australian Museum, Sydney, Australia; Dr.
Museum National d'Histoire Naturelle, Paris, France; Dr. Michelle Hamer, University
South Africa; Dr. Tsuyoshi Hosoya, The National Museum of Nature and Science, T

1

2010 GBIF Task Group: Global Strategy and Action Plan for the Digitisation of Natural History Collections



Biodiversity Informatics, 7, 2010, pp. 120 – 129.

USING GEOGRAPHICAL AND TAXONOMIC METADATA TO SET PRIORITIES IN SPECIMEN DIGITIZATION

WALTER G. BERENDSOHN AND PEGGY SELTMANN

Dept. of Biodiversity Informatics and Laboratories, Botanic Garden und Botanical Museum Berlin-Dahlem, Freie Universität Berlin, Königin-Luise-Straße 6-8, D-14195 Berlin, Germany
Correspondence e-mail: w.berendsohn@bgbm.org

Abstract. – Digitizing the information associated with by specimens in natural history collec
endeavor providing falsifiable information about past and present biodiversity
application in a variety of research fields far beyond the current applicati
digitization efforts are driven by individual institutional necessities
This has led to an over-all information resource that is patc
as in quality. Digitizing all specimens is not an
biodiversity studies are both taxonomically and geographically restricted, but access to non-digitized collection information is almost exclusively by taxon name. Creating a "Geotaxonomic Index" providing metadata on the number of specimens from a specific geographic region belonging to a specific higher taxonomic category may provide a means to attract the attention of researchers and governments towards relevant non-digitized holdings of the collections and set priorities for their digitization according to the needs of information users outside the taxonomic community.

Key words. – Natural history collections; collections; specimens; specimen data; metadata; digitization; GBIF; biodiversity research.

INTRODUCTION

Each specimen in natural history collections carries a wealth of information, most notably the past location in space and time of a verifiably identified species. Specimens add the historical component of biodiversity to contemporary observation networks, and deposited voucher specimens are reproducible scientific evidence for species identifications for all areas of biodiversity research.

General agreement exists that having all this information available in electronic form would greatly improve the information base for many research domains, including – but by no means restricted to – systematics.

Substantial efforts have been invested by institutions over the past years in digitizing specimen information, and the Global Biodiversity Information Facility (GBIF) offers the technical infrastructure to make these data records universally available. However, the GBIF Task Group on the Digitization of Natural History Collections realized that

- we are very far from a complete data resource: most specimen data remain accessible only by consulting the actual specimen
- digitizing individual specimens is a very costly process and no funding for globally comprehensive specimen digitization is in sight
- existing digitization efforts are not coordinated, producing an information resource that is patchy in taxonomic and geographic coverage as well as in quality (e.g. Yesson 2007, Balian & al. 2008, Kusber & al. 2009).
- no mechanism exists to request globally information about relevant non-digitized holdings of collection institutions. The potential of specimen information is thus not appreciated by a wider user community.

We posit that we may overcome these obstacles by making user demand the driver of detailed digitization of individual specimens. User demand for detailed specimen data comes from ongoing or projected research or is connected to

Biodiversity Informatics
2010, 7: 120 – 129

"However desirable, digitization of all specimens across the globe is a noble but impracticable goal."

"Digitizing all specimens is not an achievable aim at present"

Our collections are…

More technology, more automation, more speed

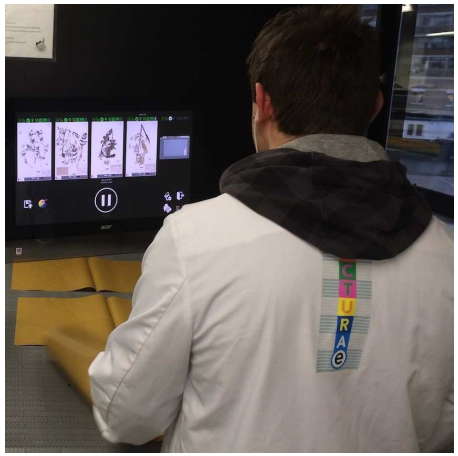
Whole drawer scanning

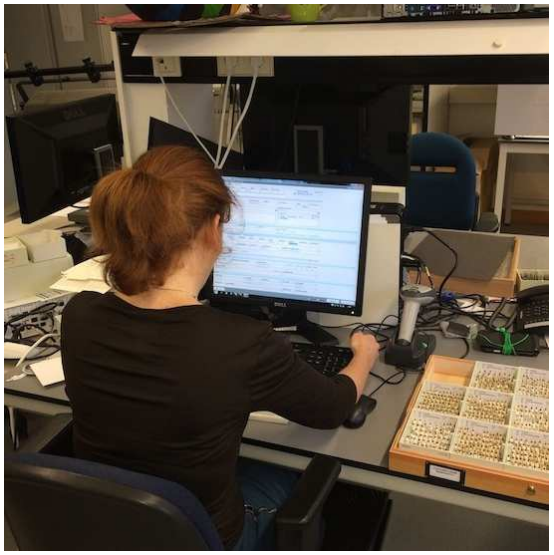
Herbarium sheet scanning

Microscope slide scanning

European collections rising to the challenge

Large-scale data capture & digitisation in France, Netherlands & Finland





NHM London Science Strategy 2013-17

A New Voyage of Discovery

Three Focal Areas

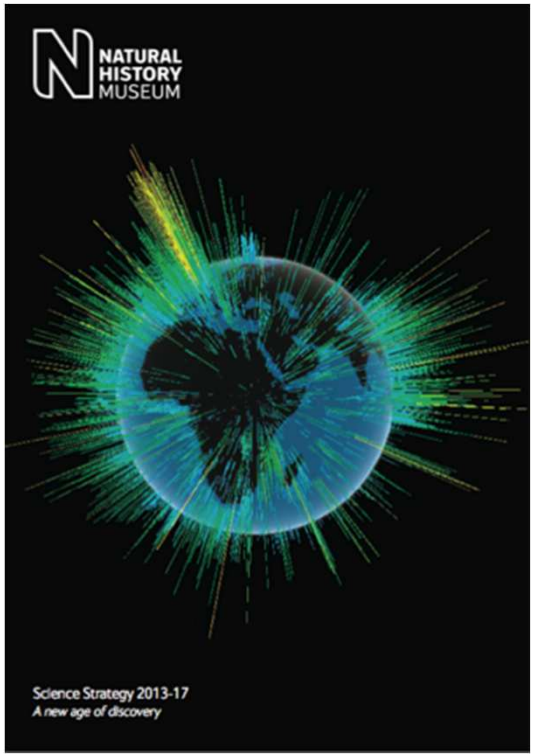
1. Scientific discovery
2. Scientific Infrastructure
3. Scientific engagement

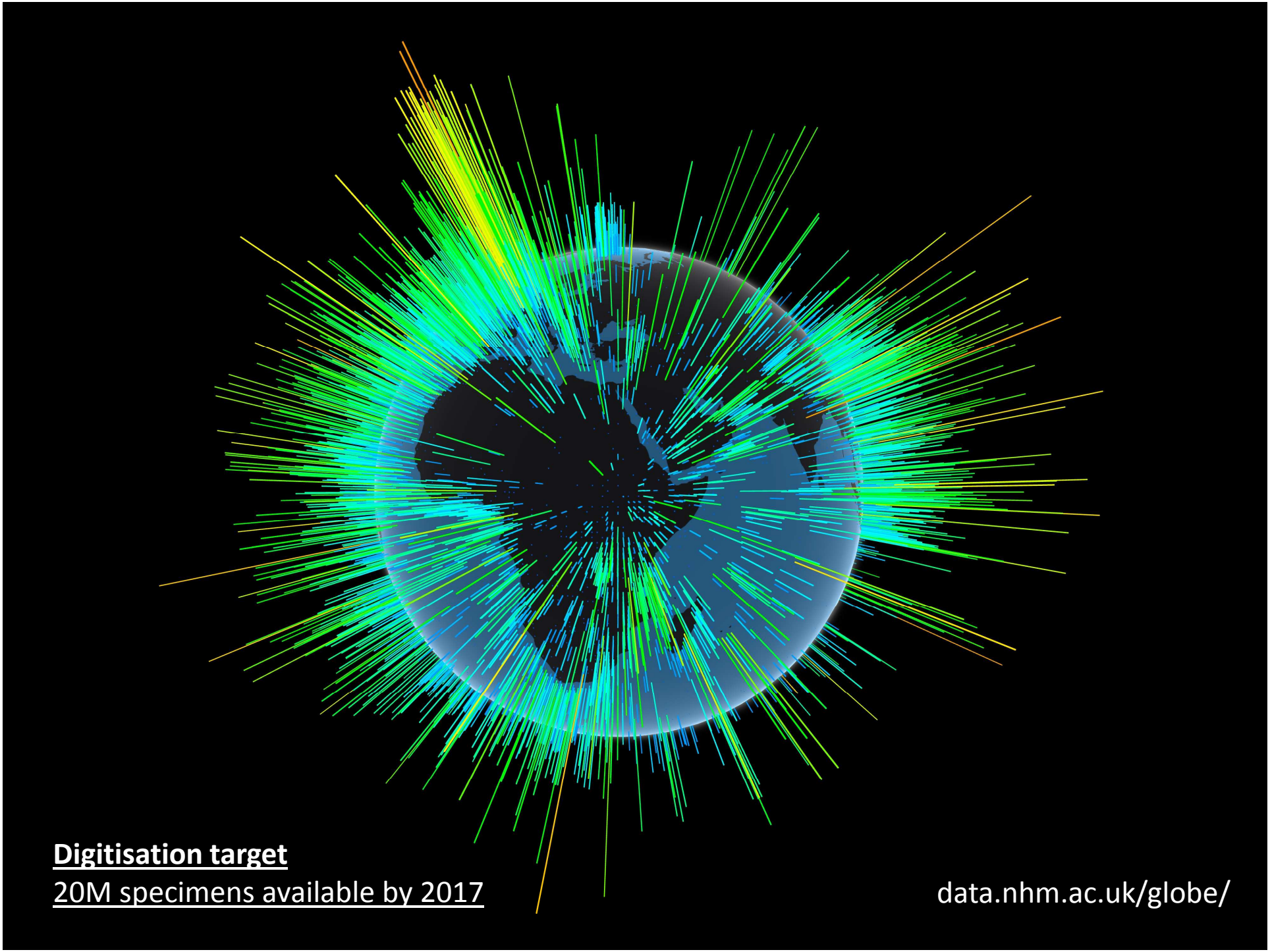
Five Challenges

1. The Digital NHM
2. Origins, evolution & futures
3. Biodiversity discovery
4. Natural resources & hazards
5. Science, society & skills

Resources & funding

Measuring success




Digitisation target
20M specimens available by 2017
data.nhm.ac.uk/globe/

A long way to go, practically, technically & culturally…

NHM collections comprise c.80m objects
Physical register: c.5m
Digital data: 2.8m
Images: 350k

Collection area	No of objects	No of type specimens	Physical register	Digital data
Palaeontology	6,919,207	43,146	2,364,232	340,636
Mineralogy	423,563	615	425,000	402,727
Botany	5,863,000	172,750	127,200	645,222
Entomology	33,753,257	612,796	57,197	255,000
Zoology	27,501,350	325,000	1,986,000	1,160,216
Library & archives	5,460,000	-	-	-
TOTAL	**79,920,377**	**1,154,307**	**4,959,629**	**2,803,801**

NHM Digital Collections Programme

A 2, 5 and 10 year plan...

To collate, organise and make available one of the world's most important natural history collections as digital resource, delivering:

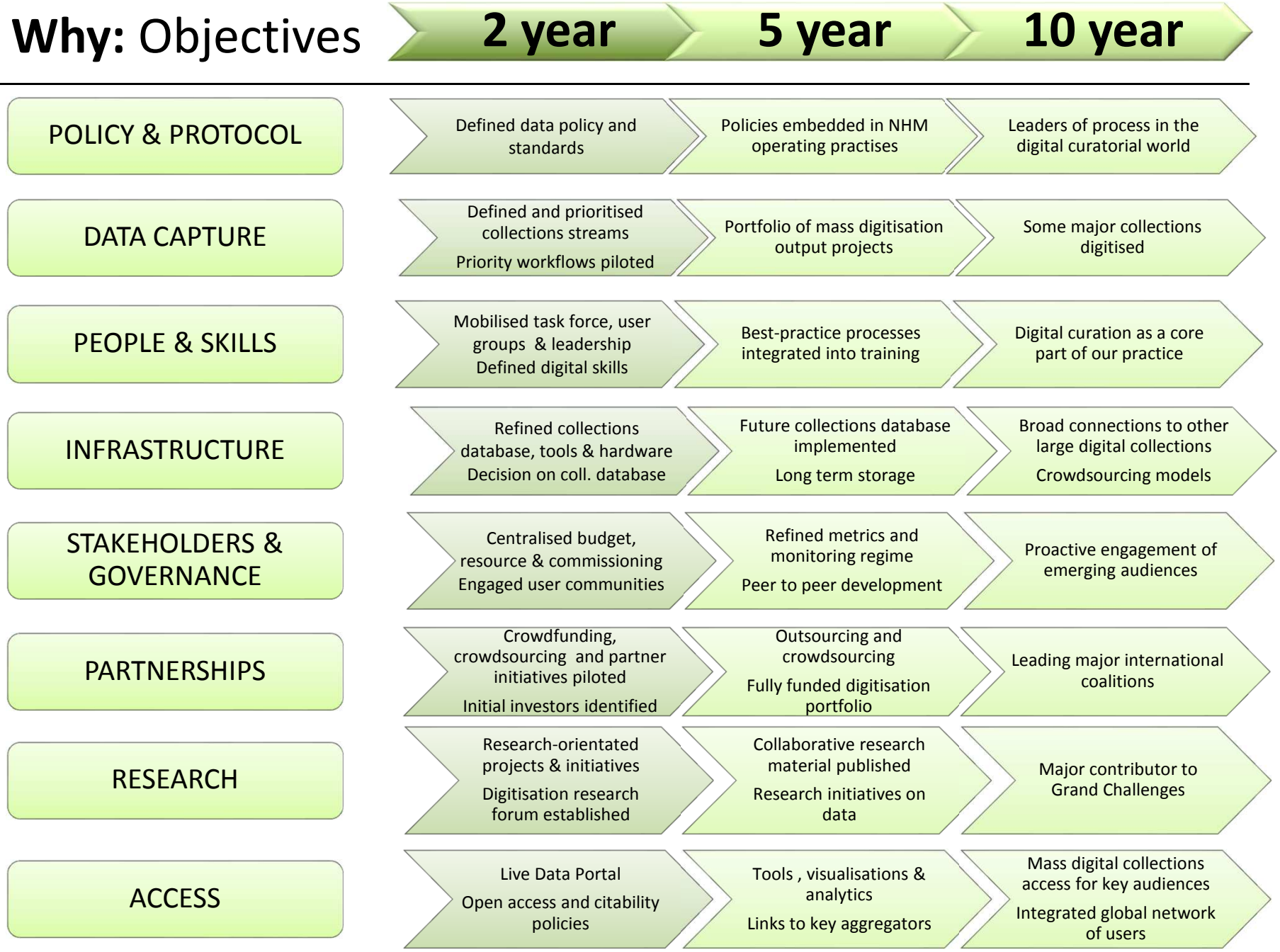
1. *an online specimen / lot-level database to manage all holdings*
2. *core meta-data and / or images for key parts of the collection*
3. *flexible informatics tools*

£750,000 for first 2 years

Outline

1. Why
   - Internal objectives & benefits
   - Research opportunity - the iCollections example
2. What
   - How much data to digitise
   - Linking digitisation effort to project benefits
3. How
   - Digi-street pilots, quick wins (herbarium, drawer & slide scanning)
   - Crowdsourcing pilots & options
4. Where
   - NHM Data Portal
   - External Portals (E.g. GBIF, Europeana)
5. Links
   - *Crowdfunding*
   - *H2020 projects (COST, SYNTHESYS, LOD, VRE, Dig. Inf.)*
   - *Other museums, herbaria & partners (e.g. CETAF & publishers)*
6. When

1. **Why:** Objectives



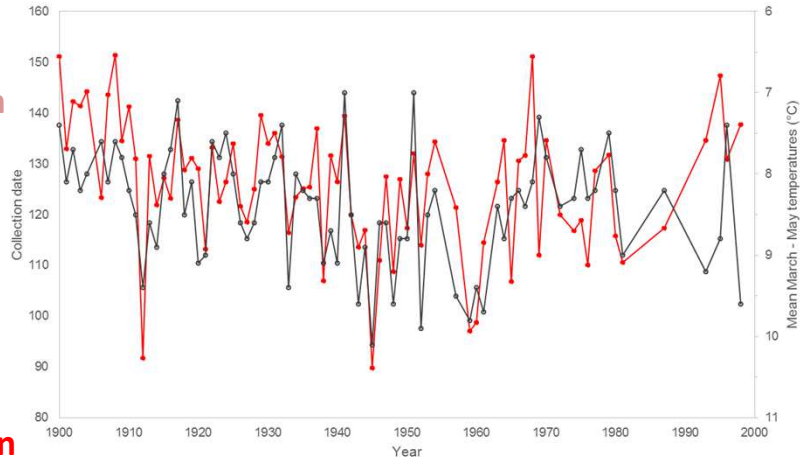
1. Why: Research opportunity & the iCollections pilot

Using the NHM collections to track long-term seasonal response of butterflies to climate change

- Digitisation of British and Irish Lepidoptera collection
- Species poor, specimen rich
- ~500,000 specimens, 5,000 drawers
- Re-curation, imaging, label data, georeferenced
- ~25% complete (started Jan.'13)
- About 50% specimens 'useable'
- Many specimens in most years (late - 19th century to 1970)
- Provide longer time perspective than most observational records (BMS post-1976)

1. Why: Research opportunity & the iCollections pilot

Later collection



Earlier collection

Cooler springs

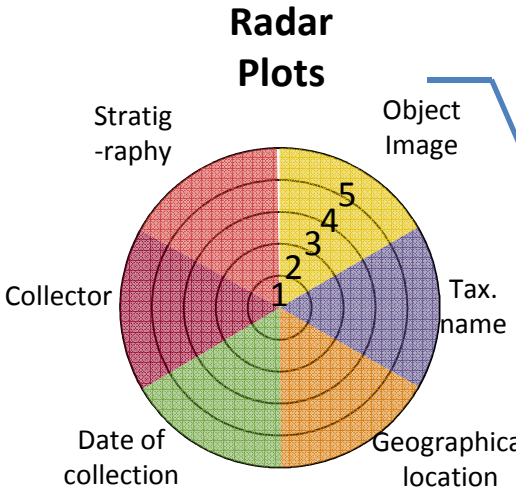
Warmer springs

Relationship between 10th percentile collection date of *Anthocharis cardamines* (Orange tip) and mean Mar. – May temp.
(N.B. temp. axis reversed)

- 1900-2000, strong correlation between initial collection dates & temperature
- Critical marker on phenological response prior to recent rapid climate change
- Longer time perspective than most observational records (BMS post-1976)
- Museum data available for rare or hard to record species
- An example of unique biological and ecological data from collections

Brooks, Self, Toloni & Sparks, 2014, *Int. J. Biometeorol.*
DOI 10.1007/s00484-013-0780-6

2. What: Linking data capture effort to research benefits



First Sweep

Benefit Justified

Level of Data Capture

Low — High

		1	2	3	4	5
Administrative	BM Number of object	Unknown/ Not assigned	In label image	Transcribed	Transcribed & linked to digitised register	Transcribed & linked to digitised register
Administrative	Object-level identifier if different	Not assigned	In label image; not machine-readable	Transcribed	Machine readable barcode	Machine readable barcode
Administrative	Location within the NHM	Defined at project level	Collection	Collection + Cabinet	Collection + Cabinet + Drawer	Collection + Cabinet + Drawer + Location within drawer
Images	Object image	None	Multi-object (e.g. drawer level); not separable	Multi-object (e.g. drawer level); separable	Single object (low-res)	Single object (high-res)
Images	Label image	None	Multi-object (e.g. drawer level); human-readable	Single object; human-readable	Multi-object (e.g. drawer level); transcribed	Single object; transcribed
Object(s) metadata	Taxonomic name	Unidentified	In label image	Transcribed but above species-level	Transcribed to species-level	Name linked to current taxonomy
Object(s) metadata	Type status (if single specimen)	Unknown	In label image	Yes/No	Transcribed & specified	Transcribed & specified
Object(s) metadata	Geographical location	Unknown or in label image	Continent (TDWG level 1)	Country or region (TDWG levels 2-4)	Georeferenced named locality (e.g., town)	Coordinates based on narrative or GPS
Object(s) metadata	Date of collection	Unknown	In label image	Transcribed year	Transcribed month and year or date range	Transcribed date
Object(s) metadata	Collector	Unknown	In label image	Transcribed	Transcribed & linked to controlled list	Transcribed; linked to controlled list & collector's notes
Object(s) metadata	Stratigraphy	Unknown	In label image	Transcribed verbatim	Chronostratigraphic interpretation	Chronostratigraphy, biostratigraphy and lithostratigraphy
Object(s) metadata	Additional data (e.g. host)	Unknown	Imaged but not transcribed	Partially transcribed	Transcribed in full	Transcribed in full and integrated

Research Areas

We have mapped out much data we need to address these questions

Futures

- Coarse analysis of spp. distribution change
- Coarse Species Distribution Models
- Phenological change

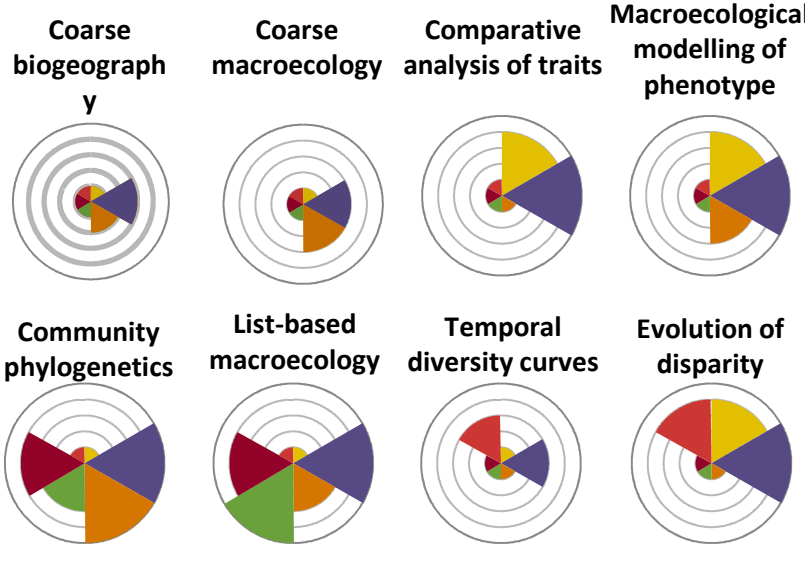
Resources

- Semi-automated capture of trait data
- Modelling within-spp variation across range

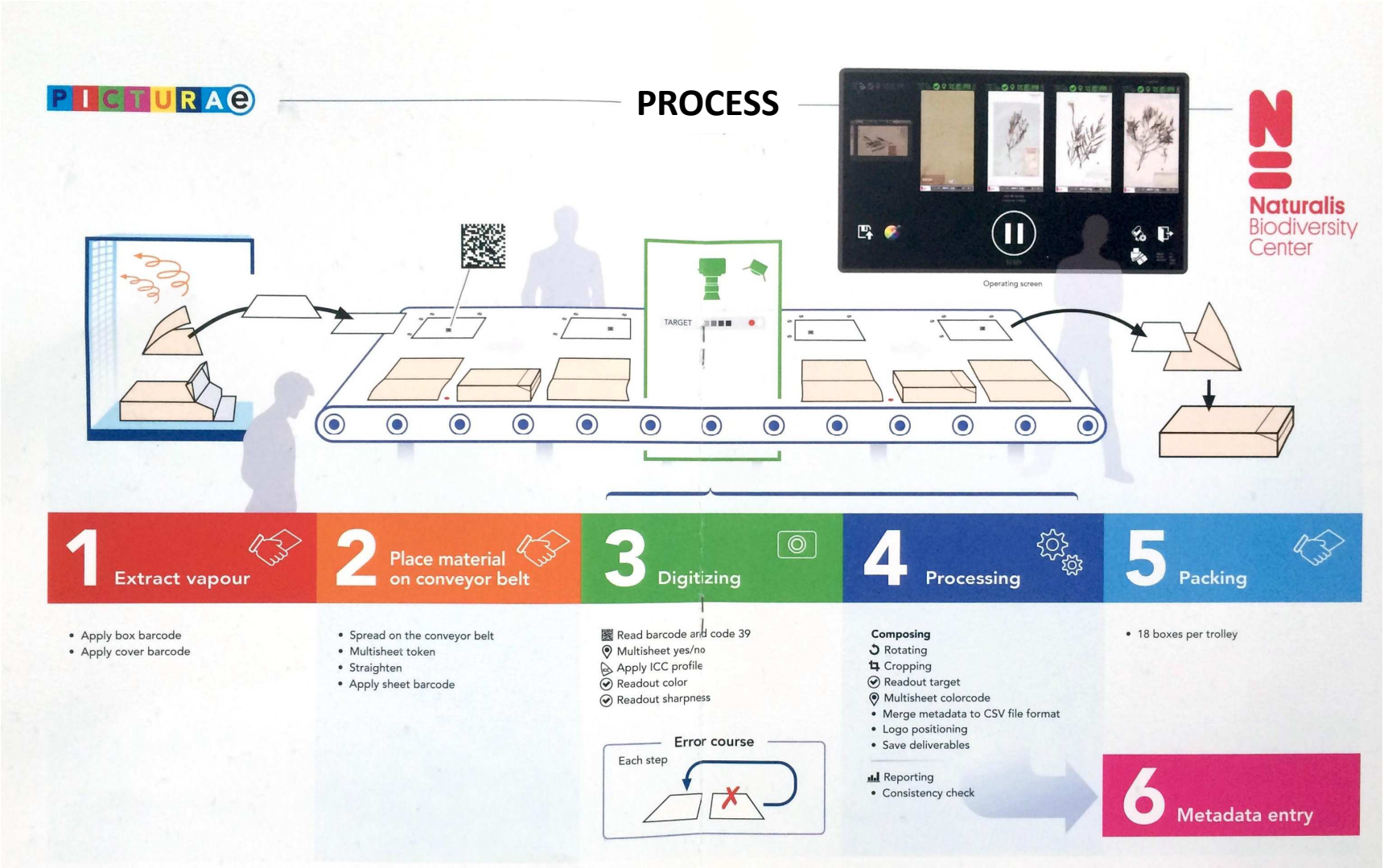
Hazards

- Earliest records of invaders
- Effects of decadal climate oscillations
- Modelling biotic consequences of weather
- Evolution of invasive species

Origins



3. How: Digi-street pilots (Herbarium Sheets)

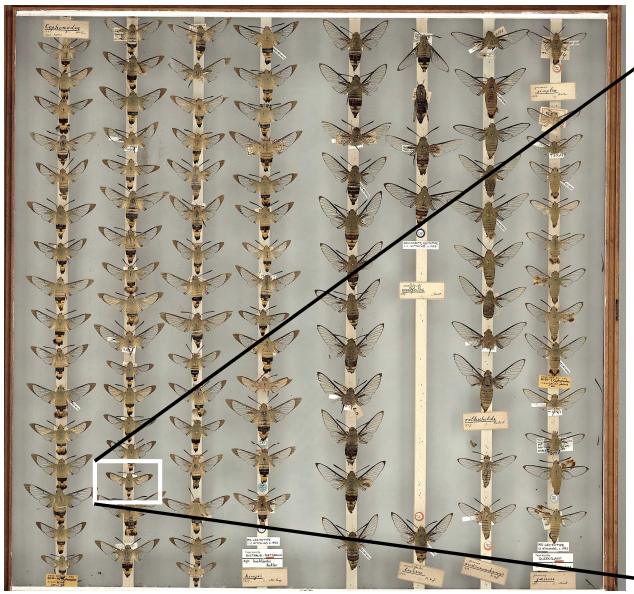


3. How: Digi-street pilots (Herbarium Sheets)

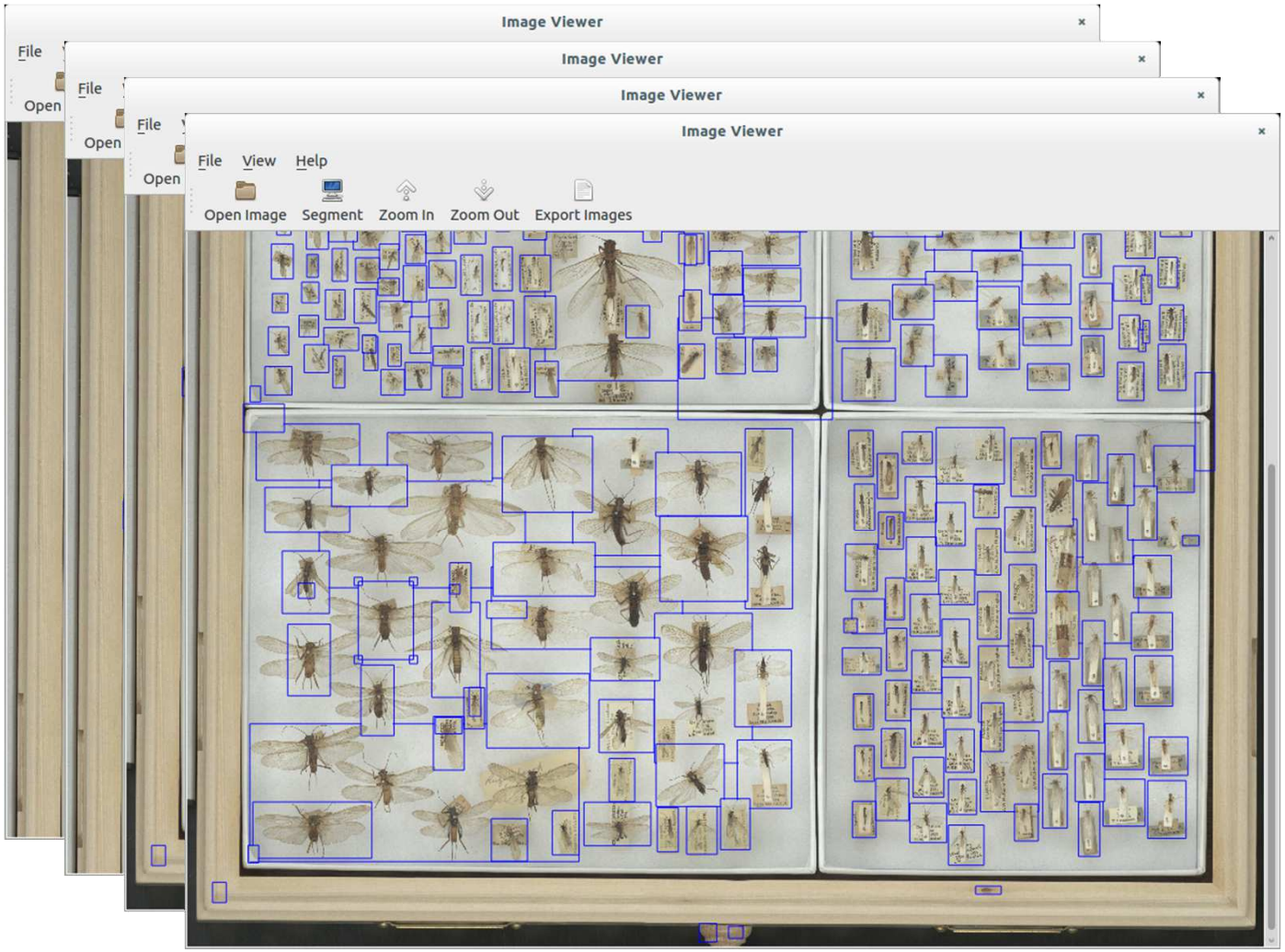
33k Specimens per day, 3 shifts (6am-10pm), Netherlands collection complete in 1.5 years
€1.29 Euros per specimen image (if outsourced), transcription at similar cost

3. How: Digi-street pilots (Drawer scanning & segmentation)

- SatScan whole drawer scanning
- 30 Million specimens, 130k drawers
- Fast, high res. multi-specimen drawer images (5 mins. each)
- No specimen handling
- Limited drawer / unit tray metadata, plus identifiers
- Specimen segmentation problem
- Digital and physical collection gets out of sync
- Need to automate specimen segmentation



3. How: Digi-street pilots (Drawer scanning & segmentation)



Starting image

↓

Auto-segment

↓

Mark errors

↓

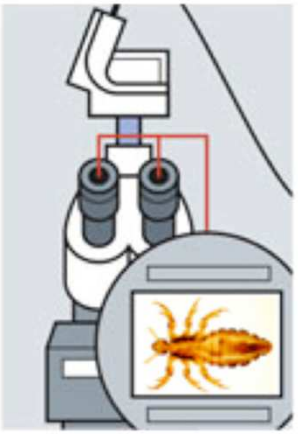
Correct

Work with Pieter Holtzhausen and Stéfan van der Walt (Stellenbosch University)
Software: Inselect
Main language: Python

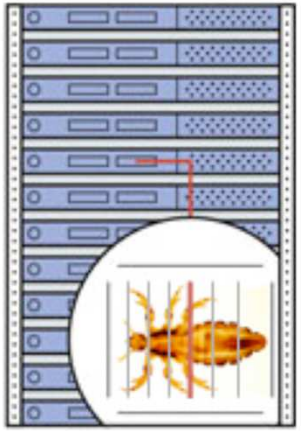
3. How: Digi-street pilots (Slide scanning)

1. Slides cleaned & barcoded

2. Loaded into hopper (50-100)



3. High resolution scan

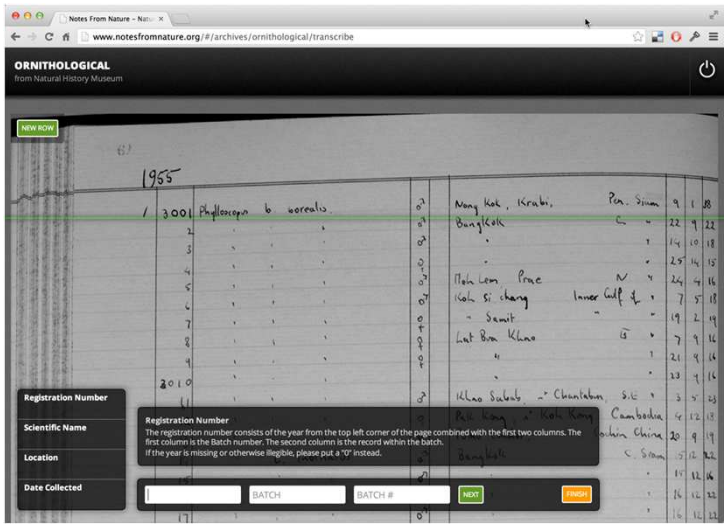


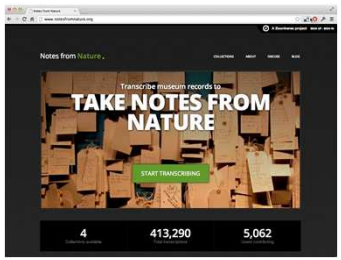
4. Images stored & databased

- Originally for histological sections
- Can be adapted for NH specimens
- Many issues:
  - Speed (Max. 500 per day)
  - File size (2-5GB per slide
  - Network ingestion (100MBps)
  - Reading labels at both ends

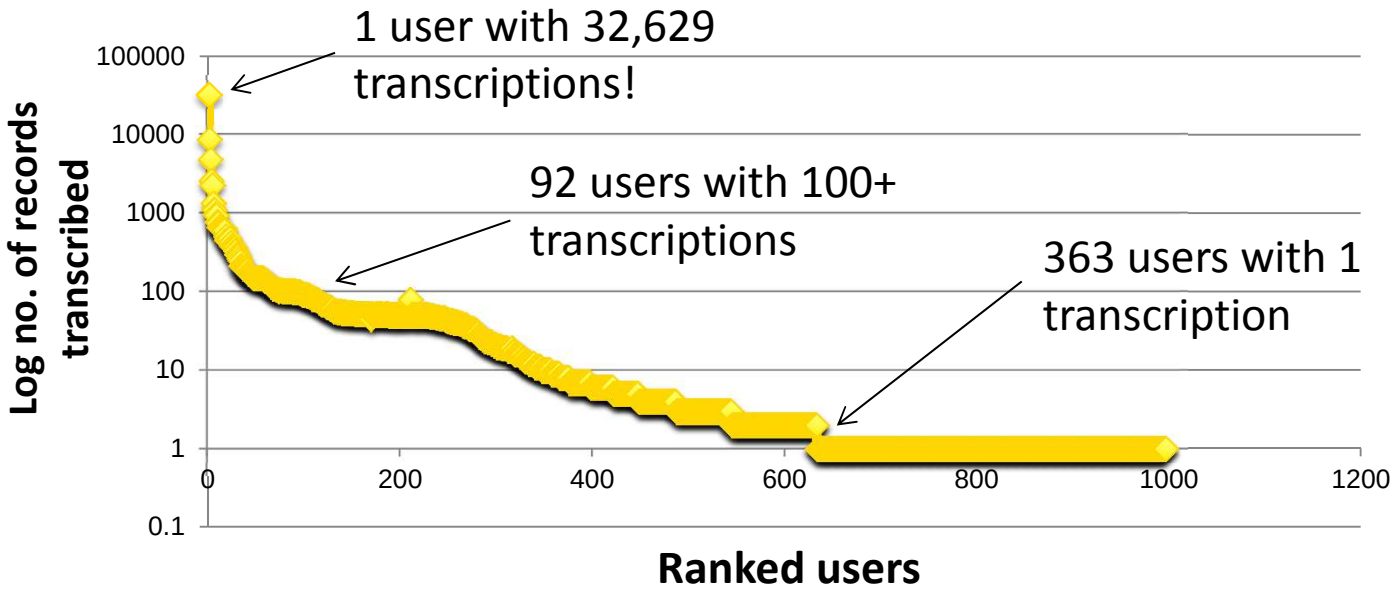
- NHM testing 6 systems
- NERC capital grant awarded
- Fully operation early 2015

3. How: Crowdsourcing pilot

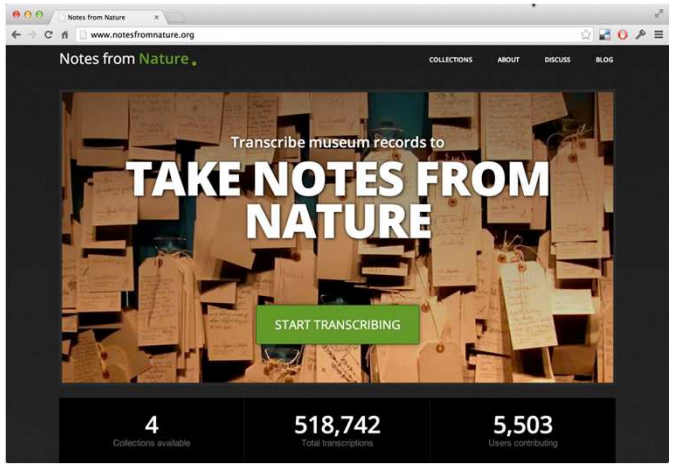




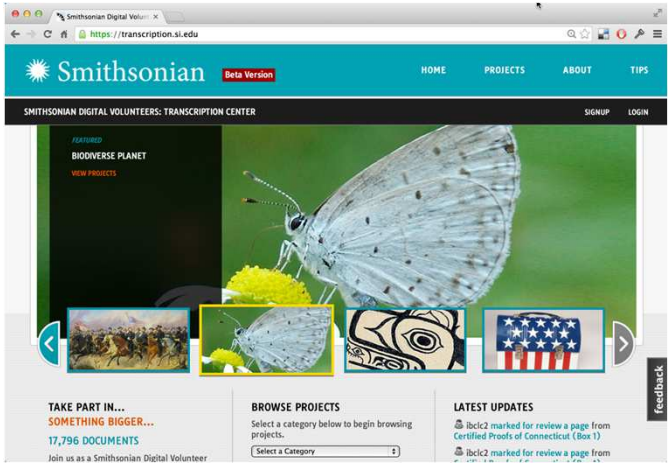
- NHM Bird registers
- No advertising
- Hard to transcribe
- Challenging starting project



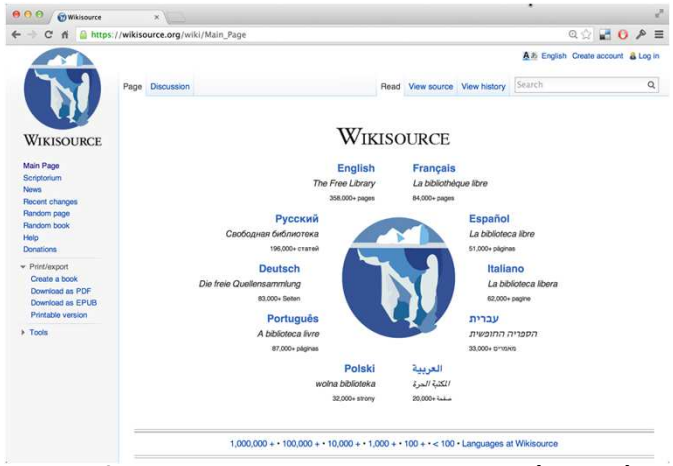
3. How: Crowdsourcing options



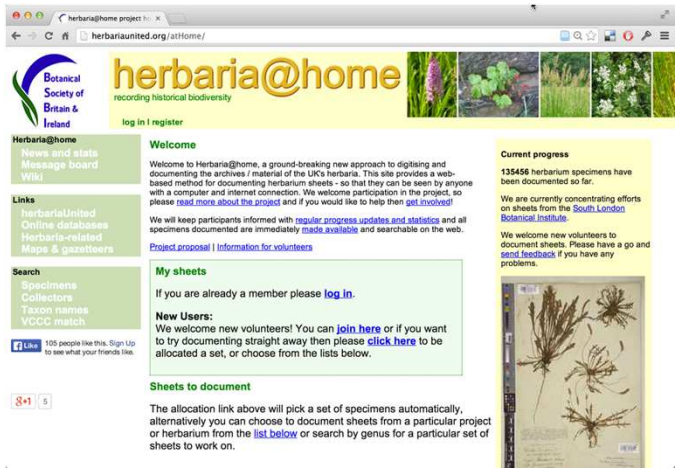
Zooniverse Projects



Smithsonian Digital Volunteers



Wikisource transcription (WiR)

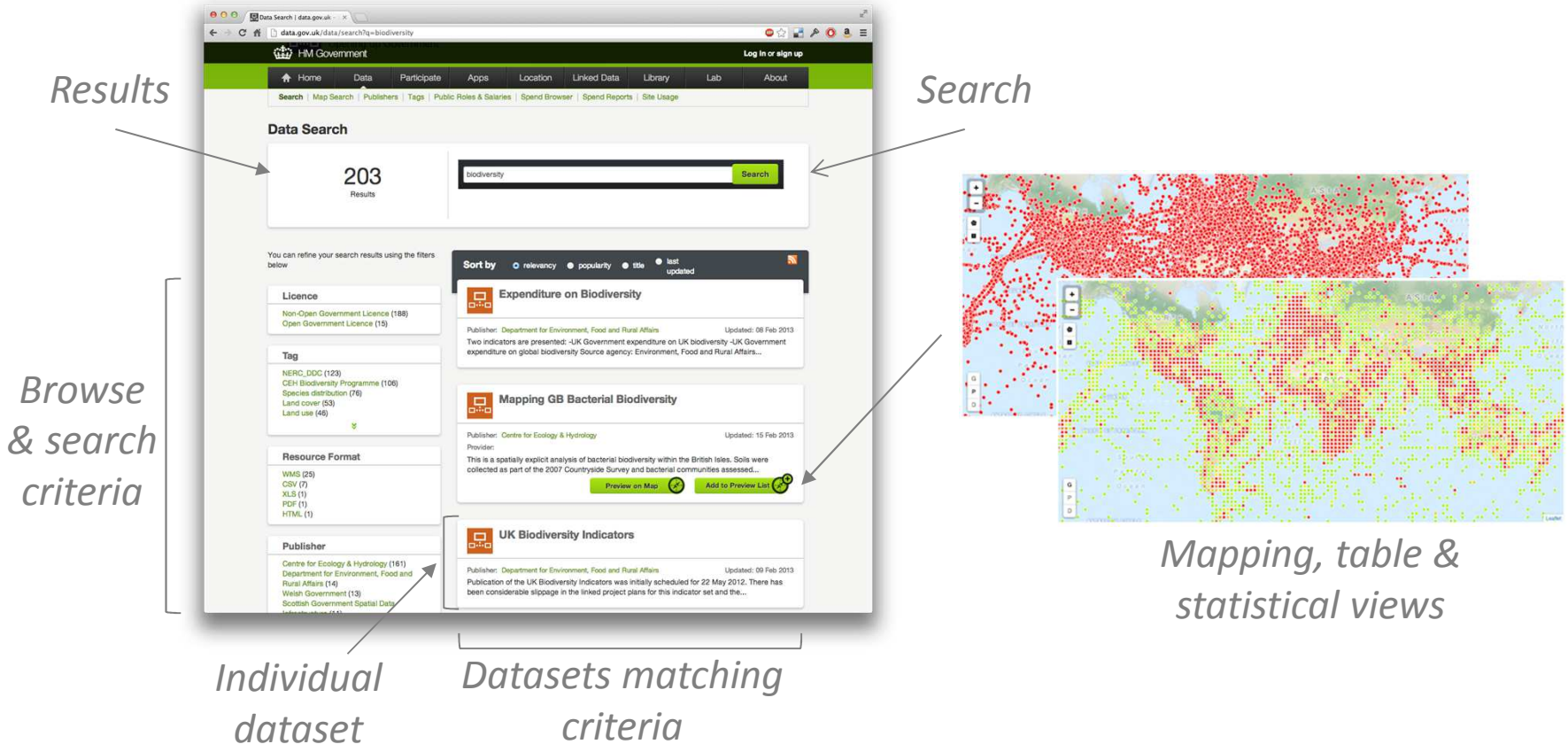


Herbarium@Home

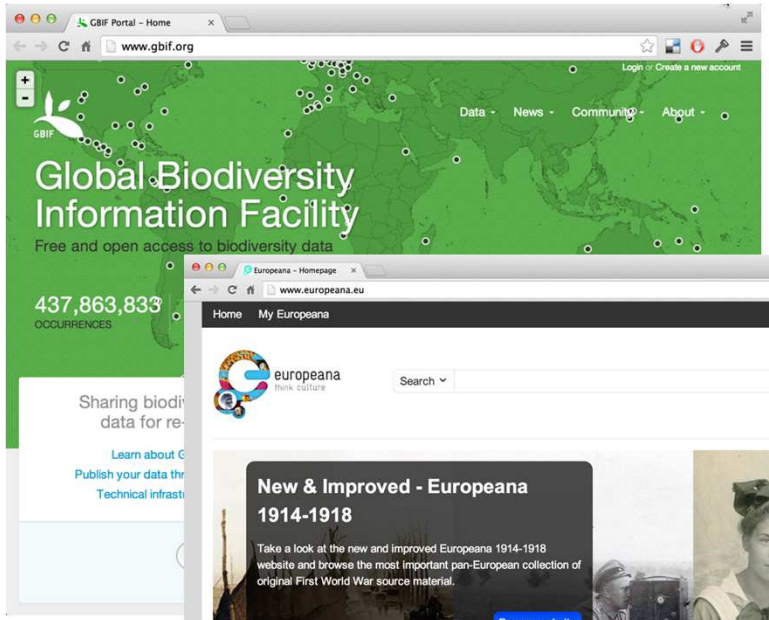
Next steps: Survey and review of natural history transcription projects cf. paying transcribers

4. **Where:** NHM Data Portal

- A focus for deposition and discovery of NHM research & collections data
- Stable, citable identifiers on datasets & specimen / lot records
- Transparent data quality (un-reviewed, reviewed, reviewed & updated)
- Download (DwCA), web-services & Linked Open Data
- Build using CKAN, with enhanced mapping functionality

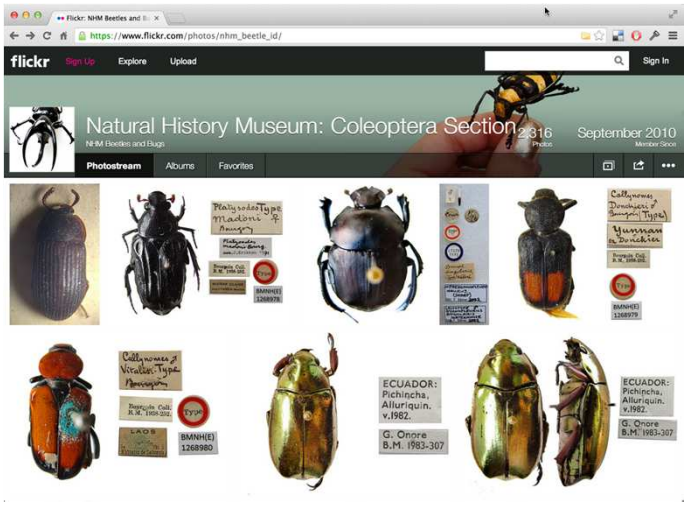


4. **Where:** External Portals

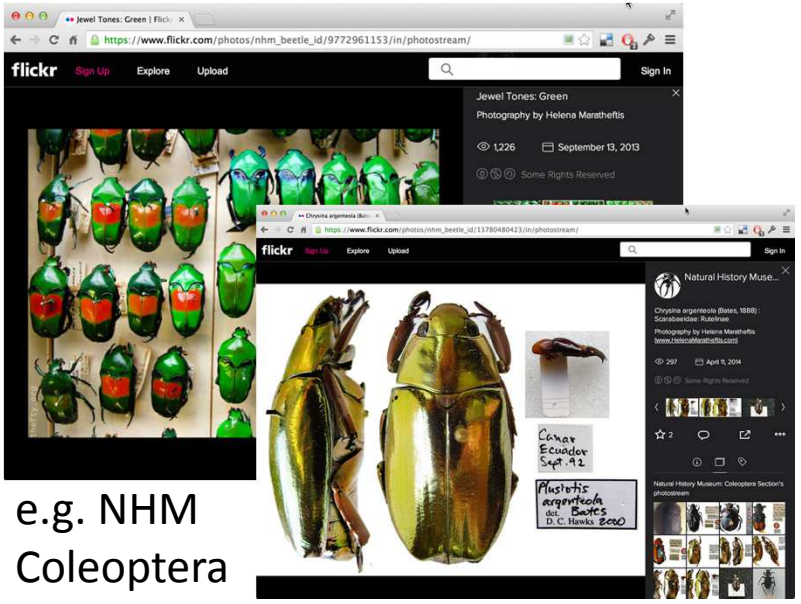


GBIF

Europeana



Flickr



e.g. NHM Coleoptera

- NHM almost getting data to GBIF!
- Submitting to Europeana portal (via Open-Up)
- Niche collections on Flickr
  - Robust API services
  - Gateway to image analysis projects (e.g. species recognition & trait extraction tools)

5. Links

Crowdfunding

- Personalizes donation
- Scales well
- Requires lots of data
- Most crowdsourcing platforms unsuitable
- Potential for a data visualization to support our needs

H2020 Projects

- EU Research & Innovation funding Programme
- €80 Billion from 2014-2020
- Strong record (EDIT, ViBRANT, SYNTHESYS1/2/3)
- 5 proposals in development for 2014/15
- Better alignment with Digital Collections Programme

Partners

- Major museums & herbaria (Kew, Smithsonian, & Euro.6)
- Umbrella organisations & projects (GBIF, CETAF, iDigBio)
- Universities (e.g. on Image analysis)
- Data publishers (engagement on data & systems)

6. When

Key dates over next 2 years

Herbarium scanning

- Pilot – TBC (starting late-2014)

Drawer scanning

- Segmentation Software (Aug. 2014)
- Pilots (Ongoing)

Slide scanner

- Testing 6 systems (Complete)
- Procurement / purchase (July 2014)
- Pilot projects & system integration (From Sept. 2014)

Crowdsourcing pilots

- Draft review paper (Aug. 2014)
- Additional Notes from Nature Project (early 2015)

NHM Data Portal

- Internal release (June 2014)
- Public release (Jan. 2015

Funding

- H2020 projects (submitted, Sept. 14 & Jan. 15)

Acknowledgements

Digital Collections Programme

Planning: Ian Owens, Ben Atkinson, Dave Thomas, Andy Purvis, Emilie Smith & Vince Smith.

iCollections

Project team: Gordon Paterson, Geoff Martin, Martin Honey, Blanca Huertas, Darrell Siebert, Vladimir Blagoderov , Steve Cafferty, Adrian Hine, Chris Sleep, Mike Sadka, Elisa Cane, Lyndsey Douglas, Joanna Durant, Gerardo Mazzetta, Flavia Toloni, Peter Wing, Malcolm Penn & Liz Duffle.

Research: Steve Brooks, Angela Self, Flavia Toloni & Tim Sparks.

Drawer scanning

NHM Satscan development: Vladimir Blagoderov, Laurence Livermore & Vince Smith.

Software: Pieter Holtzhausen & Stéfan van der Walt (Stellenbosch University).

Slide scanner

Testing: Vladimir Blagoderov & Alex Ball.

Crowdsourcing

Pilots (NHM Team): Tim Conyers, Lawrence Brooks & Adrian Hine.

Review paper: Laurence Livermore & Vince Smith.

NHM Data Portal

Project team: Vince Smith, Darrell Siebert, Dave Thomas & Adrian Hine.

Development: Ben Scott & Alice Heaton.

Apologies to anyone I have missed!